# **Checkpoint 1 - Grupo 12**

## **Análisis Exploratorio**

El dataset sobre el que vamos a trabajar es un registro de anuncios de propiedades en venta en Argentina durante el año 2021. Para empezar, tenemos un total de 460.154 registros, es decir, anuncios. Para cada anuncio contamos con 20 columnas.

En esta primera etapa vamos a trabajar sobre qué columnas son importantes y de cuáles podemos prescindir. Para empezar, vamos a reducir nuestro estudio a las propiedades de tipo PH, departamento o casa dentro de Capital Federal cuyo tipo de operación sea venta en moneda USD. De esta forma la cantidad de registros válidos con los que nos quedamos son 89.709.

**Preprocesamiento de Datos**

Detallar las tareas más importantes que realizaron sobre el dataset, les dejamos algunas preguntas cómo guía:

## ¿Se eliminaron columnas (Nombre de la columna y motivo de eliminación?

1. ¿Detectaron correlaciones interesantes (entre qué variables y qué coeficiente)?
2. ¿Generaron nuevos features?

## ¿Encontraron valores atípicos?¿Cuáles?¿Qué técnicas utilizaron y qué decisiones tomaron?

## ¿Qué columnas tenían datos faltantes?

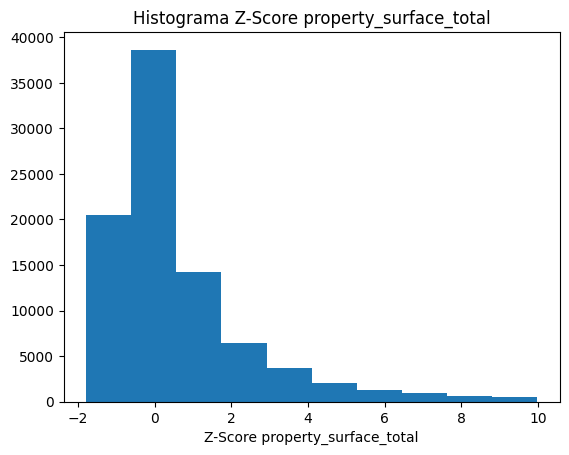
¿En qué proporción? ¿Qué se hizo con estos registros?

Luego de la filtración según los parámetros establecidos, vamos a ver los datos nulos dentro del dataset bajo estudio. Cuando realizamos el porcentaje de celdas vacías dentro de todas las columnas, observamos que las columnas *place\_l5 y place\_l6* tienen un 100% de celdas nulas. Sin duda, eliminamos dichas columnas. Ahora, *place\_l4* tiene un porcentaje de nulos del 96%, valor muy elevado como para hacer algún tipo de corrección o relleno, por lo tanto, eliminamos la columna. De esta forma eliminamos directamente tres columnas.

Luego nos encontramos con un porcentaje mucho más cercano a cero, pero de variables que están relacionadas entre sí y que además son muy importantes para el dataset. Estas columnas son *property\_surface\_total* y *property\_surface\_covered* que tienen un porcentaje de nulos de 5% y 3.5% respectivamente. El método que vamos a utilizar para rellenar las celdas vacías va a ser el de rellenar los datos a partir de la media de dichas columnas y el cociente entre ellas, de ahora en más denominado **cociente**. Antes de seguir, chequeamos que ninguna propiedad tenga ambos valores nulos y para cada celda vacía de la columna *property\_surface\_total* vamos a multiplicar su valor de *property\_surface\_covered* por el cociente. De forma análoga, hacemos lo mismo para los valores nulos de *property\_surface\_covered*, pero esta vez, dividiremos por el cociente el valor de *property\_surface\_total*

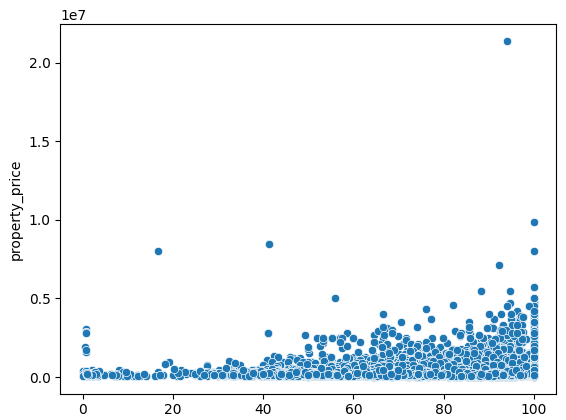
Más adelante en nuestro análisis, vemos que la columna de *id* no va a ser nunca utilizada y que la columna *created\_on* es redundante ya que la columna *start\_date* brinda la misma información.

A la hora de las visualizaciones fuimos encontrando valores que carecen de sentido, como propiedades con decenas de miles o centenas de miles de metros cuadrados de superficie. Dichas anomalías se detectaron fácilmente con un dispersograma para luego confirmar de manera más eficiente con el método de z-score que observamos en el siguiente gráfico:



**Visualizaciones**

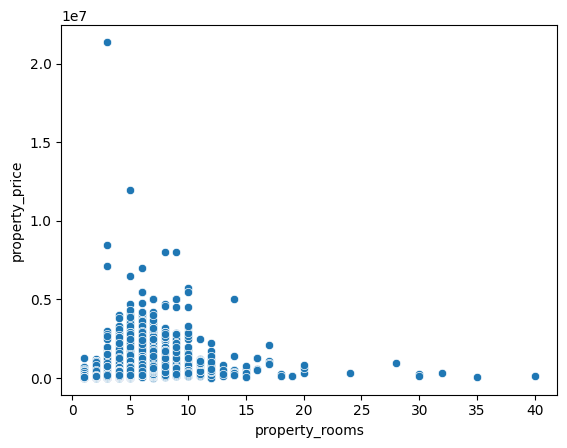
1. **Precio de propiedad VS Porcentaje de superficie cubierta**



En este gráfico se puede ver el precio de propiedad (en el eje Y) comparado con el porcentaje de superficie cubierta (en el eje X). Este porcentaje se calcula con la siguiente fórmula:

Nos pareció interesante ya que la relación entre superficie cubierta y no cubierta parece tener un muy ligero crecimiento cuadrático, lo cual puede ser útil para predecir precios.

1. **Precio de propiedad VS Porcentaje de superficie cubierta**

****

En este gráfico se puede ver el precio de una propiedad (en el eje Y) comparado con la cantidad de ambientes de la misma (en el eje X)

Nos pareció interesante ya que, si bien uno pensaría que cuanto más ambientes tiene una propiedad mayor el precio, este gráfico sugiere que la relación tiene una forma de campana de Gauss. Por lo tanto, este gráfico deja en evidencia que la cantidad de ambientes quizá no sea la mejor variable para predecir precios cuando la misma tiene valores altos.

**Clustering**

Para determinar la cantidad de cantidad apropiada de grupos se usa el método de codo.

Para esto se usa el algoritmo de Kmeans con un número de clusters (k) entre 2 y 10.

Se graficó el SSE en función de k y se buscó el punto de inflexión, dicho punto se da en k=3.

También se graficaron la longitud y latitud en un mapa de la CABA, coloreados por cluster.

**Estado de Avance**

1. **Análisis Exploratorio y Preprocesamiento de Datos**

**Porcentaje de Avance:** 50%/100%

**Tareas en curso:**

* Desarrollar la sección incompleta de valores atípicos
* Resolviendo las propiedades que tienen los dormitorios y habitaciones en Nan

**Tareas planificadas:**

* Pulir las visualizaciones ya realizadas, y rehacer las que no proporcionan información al lector
* Averiguar si podemos analizar correlación entre la variable que no sean la de Pearson
* Qué hacer con las propiedades que no tiene un barrio asignado
* Comparar las distribuciones cuando hago un input

**Impedimentos:**

* Utilizar apropiadamente los frameworks de visualizaciones para hacer visualizaciones más limpias e informativas
* El Z-score Modificado de Superficie Total dió muchos outliers (más de 7800) y no sabemos qué hacer con ellos (ya que si los eliminamos corremos el riesgo de recortar muchas entradas del Dataset)
* Datos faltantes y Outliders, ya que se puede trabajar de manera certera, debido que esta falta de datos o Outlander puede afectar de manera directa al resultado esperado: Por ejemplo en un momento se buscaba la correlación de Pearson entre el total de terreno y la cantidad de habitaciones, y había una propiedad que decia que tenia un terreno de 50000m2 con 2 habitaciones.

1. Exploración Inicial:
   * Deberia tener mas grafico univariables, por que siento que no hay suficientes
2. Visualización de los datos: Crear más visualizaciones que relacionen 3 variables, por ejemplo:
   * Precio por metro cuadrado vs Barrio
   * Precio por ambiente vs Barrio (o Precio vs Habitaciones vs Barrio (con un heatmap))
   * start\_date/end\_date vs Precio (Para ver si el período donde se publicó afecta al precio)
3. Datos Faltantes:
   * Hay pocos datos que siguen estando en Nan, hay que analizar qué hacer con ellos, serían los atributos habitaciones, domitoria y barrio los que siguen teniendo datos nulos
4. Valores atípicos:
   * Encontrar outliers multivariados, como por ejemplo:
     + Ambientes + Dormitorios (Descartando los que tienen Ambientes < Dormitorios)
     + Precio + Proporción entre Superficie Total y Superficie Cubierta (Como se vió en la sección de Visualizaciones)
     + Precio + Superficie (Pedida por el TP)
   * Describir y decidir el tratamiento de los outliers

**2. Agrupamiento**

**Porcentaje de Avance:** 65%/100%

**Tareas en curso:**

* Graficando la longitud y latitud en un mapa de la CABA con coloreados por cluster.

**Tareas planificadas:**

* Intento entender en función de qué características fueron formados los grupos.

**Impedimentos:**

* En función de qué características fueron formados los grupos.

## **Tiempo dedicado**

Indicar brevemente en qué tarea trabajó cada integrante del equipo durante estas semanas.Si trabajaron en las mismas tareas lo detallan en cada caso (como en el ejemplo el armado de reporte). Deben indicar el promedio de horas semanales que dedicaron al trabajo práctico.

| Integrante | Tarea | Prom. Hs Semana |
| --- | --- | --- |
| Ramiro Gareis | Exploración Inicial  Datos Faltantes | 4 |
| Elvis Claros | Clustering | 3 |
| Matías Venglar | Visualizaciones  Detección de Outliers | 3 |
| Gonzalo Olmos | Armado de Reporte | 1 (Gonzalo Olmos se incorporó al grupo 2 días antes de la entrega del checkpoint 1) |